Ontology-based Data Exchange and Integration: an Experience in CyberInfrastructure of Sensor Network Based Monitoring System

Chen-Chieh Feng, Liang Yu
Department of Geography
National University of Singapore
1 Arts Link, Singapore 117570
{geofcc, geoly}@nus.edu.sg

***Abstract* — Scientific research has become interdisciplinary and collaborative, of which sharing and utilizing data in an efficient manner is critical. Data collected for environmental monitoring and modeling, however, often lack semantic information vital for efficient data sharing, thereby causing semantic gaps between the data collection and utilization. The problem is especially acute when data have to be processed without human intervention. To support efficient data sharing, this paper proposes an ontology-based architecture to integrate heterogeneous data. With the help of ontology reasoning, it provides a simpler and more intelligent way for data searching with high-precision and high-recall.**

Keywords - Sensor Network; Ontology Reasoning; Alignment

I. INTRODUCTION

Scientific data are collected and exchanged across many research groups. As the volume of data and range of applications increase, being able to access the data that suit our needs has become a demanding process. Although we could easily access any existing data resources through the Internet, it is often difficult to utilize them due to various heterogeneities between different data sources. Semantic heterogeneity, in particular, presents a major problem for data integration in any interdisciplinary projects [1, 2]. Such problem exists because researchers from different disciplines commit to different domain knowledge and vocabularies, thereby generating semantic gaps that must be bridged before data from different research groups can be integrated.

To cope with this problem, an information system needs to be able to parse and analyze intelligently the search issued by researchers – it should understand user queries, automatically identify data with compatible semantics, and return the data to the researchers. Over years, various tools and technologies have been explored to achieve this goal. Metadata systems have been used to assist determining the usability of datasets [3]. Such metadata systems often work for single domains. Their capability to support data queries in an interdisciplinary project is therefore not guaranteed. Ontology technology, on the other hand, has been introduced to make explicit data semantics and to support data integration at the semantic level. Ontology-driven infrastructure has increasingly gained recognitions [1, 4, 5], especially in spatial science [6, 7]. Several reusable upper ontologies have been developed [8-10]. They provide foundation for developing domain ontologies that can be easily integrated. All these developments suggest that ontology technology is a promising tool to bridge the semantic gaps facing an interdisciplinary project.

The paper aims to develop an ontology-based infrastructure to support semantic data access in an interdisciplinary project involving pervasive monitoring and modeling of the physical environment using sensor-network. Ontology is an explicit specification of conceptualization that encodes inter-connected concepts [11]. It can be extended easily to accommodate an unlimited number of concepts compared to traditional metadata systems. Its capability to support reasoning is extremely advantageous for bridging the semantic gaps as no additional classification or annotation is needed. We developed an ontology and related functions as the core components of a cyberinfrastructure for the sensor network, aiming at helping users from different domains utilize the sensor data efficiently. It has the following components: (1) a user interface that accepts queries from the users and returns query results to the users, (2) a reasoning engine that supports intelligent search, and (3) validators that verify metadata and data formats. The system is distributed and the data in the system are managed in separate databases, each of which stores data developed or processed by a research group. Each group uses a unique set of concepts and constraints to describe the meanings of its datasets. The concepts used to query the data are domain-specific.

The paper is organized as follows. The next section introduces the related work and our approach. Section III discusses our ontology design. Section IV introduces how ontology reasoning is used to facilitate data integration. Section V demonstrates the ontology alignment technique and how to connect the raw data with our ontology. Section VI presents the current system implementation. Section VII provides conclusion and presents future work.

II. RELATED WORK AND OUR APPROACH

Over years ontologies that can potentially be used to support data search and integration have been developed. Upper ontologies such as BFO [10], DOLCE [8], and SUMO [9], were developed as foundation ontologies on which various domain ontologies can be developed and then be used to facilitate data search and integration. SWEET (Semantic Web for Earth and Environmental Terminology) [12], a comprehensive ontology for the earth science domain,

 ISBN: 978-1-61208-104-5

has been used in various scientific projects [13, 14]. Its main concepts such as *Data*, *PhysicalProperty*, *Substance*, are critical for describing data semantics. However, it has few relations between concepts, which are essential for reasoning between concepts [15].

The use of ontology to search and integrate data has greatly improved the search result [1]. For example, Couchot [4] used a minimum set of concepts, which he called it reduced ontology, to build up descriptive graphs to summarize the content of the web resources. With fewer constraints than a classical ontology, the reduced ontology is more flexible and easy to use. Shah et al. [5] used ontology to annotate biomedical databases so that the data in the databases can be located with ontology concepts.

Many ontology-based methods have been proposed for data integration within distributed data infrastructures. Beran and Piasecki [16] presented a ontology-driven design for an integrated water data system based on SWEET and GCMD (NASA's Global Change Master Directory). To improve both the recall and precision for data searching in different granularity, they proposed a four-layer ontology: navigation, compound, core and detail, each represents a different abstraction level. The navigation layer contains higher-level concepts that make it easy to visualize the ontology. The compound and core layers contain concepts for assisting users' input. The detail layer contains finer concepts of those in the core layer. These concepts are used during search and for clustering the search result.

Ludäscher et al. [17] proposed a multiple-tier mediation framework for integrating data from different types of data formats, such as database and XML file. The framework aims to alleviate data users from coping with various data formats. They introduced a conceptual model wrapper layer (GM-Wrapper) that encapsulates the methods to access data directly and a generic conceptual model (GCM) layer to which the data access methods are mapped. The GCM is then mapped to an integrated view that provides easy data access for the users.

Based on the degree of efficiency and flexibility, Wache et al. [18] classified ontology-driven data integration approaches into single ontology, multiple ontology, and hybrid ontology approaches. The single ontology approach is efficient; the multiple ontology approach is flexible; the hybrid ontology approach achieves both and is thus the preferred method. Buccella et al. [19] evaluated several well-known geographic data integration systems. The result of their work suggests that most such systems are now ontology-based, but the level the geographic information represented, the degree formal representation of ontologies adopted, and the criteria used to determine how integration should proceed vary from one system to another. They thus recommend full inclusion of geographic information into the integration process, a wider adoption of formal model for ontology representation, and a better assimilation of the geographic knowledge (e.g., quantitative and qualitative relations and scale) in the integration process.

Many annotation schemes have been proposed to tag meaning to the data generated by sensors and to improve the efficiency of data exchange. Russomanno et al. [20] developed OntoSensor, a sensor ontology based on SUMO, SensorML [21] and ISO 19115, to define schema required for geographic information and services. It provided a solid conceptual foundation for sensor itself, but lacks certain concepts related to data processing, e.g., calibration, unit, process chain, and input and output. To bridge the semantic gap between sensor data and to solve the disagreement on methods for data access and exchange, Shankar et al. [22] compared the difference between the adoption of a bottom-up, entity-oriented schema construction approach and a top-down, ontology-based approach in creating a conceptual schema for integrating data generated in a wide area sensor network. They argued that the top-down approach provides semantic commonality and enables better implementation interoperability if adhering to an advertised vocabulary, thereby a higher level of semantic interoperability, is the priority for the system design.

The review shows that various ontology designs for facilitating data sharing have been examined and evaluated. To complete our system, however, more work is needed. To be more specific, we need to perform the following four tasks:

Task 1. Generate requests. A user specifies the query criteria, which includes the theme (e.g., rainfall and temperature) and the constraints (e.g., year, spatial domain, and value ranges). The user does not know the availability of the data that are related to these data types and how they are specified in the system.

Task 2. Parse and analyze requests. The reasoning engine translates the user query to an ontology query using the semantic rules defined in ontology.

Task 3. Retrieve Data. This is the process in which the computer system locates and queries heterogeneous data sources, identify suitable data, and then integrates these data into a usable format. The process relies on the alignment between ontology and data schema.

Task 4. Data production and publishing. Data are original generated by sensors and then processed and reorganized. A dataset needs to be registered and aligned to ontology before it becomes searchable and amenable to integration.

These tasks reflect the need of mediating communication between data provider's and user's sides. Task 1 is different from a traditional concept searching for that the query criteria needs to be made, which concerns the data model and numeric representation rather than a usual domain semantics. Task 2 requires a process to convert the query with the help of ontology reasoning. Task 3 requires the alignment between ontology and different types of data sources. Task 4 requires the semantics to be transferred from the data providers to cyberinfrastructure, for which we need to use the existing metadata to populate the semantics defined in our ontology.

To accomplish these tasks we developed our ontology based on SWEET and complimented it with terms from CSDGM (Content Standard for Digital Geospatial Metadata) [23] and SensorML [21] as they are standards for describing semantics of spatial datasets and sensor systems, respectively. We also explored how our ontology can be

 ISBN: 978-1-61208-104-5

used with existing metadata systems, which provides valuable information to populate its concepts and then be used for searching and reasoning.

III. ONTOLOGY DESIGN

The domains dealt with in this paper mainly include the monitoring and modeling of urban airshed and ocean water quality. For the first domain, attention has been paid to measure various attributes of air (e.g., temperature and humidity). The characteristics of the buildings that constitute urban canyons, specifically their facades, shapes, and functions, as well as the gaps or holes, such as roads and green spaces in between buildings, are also important concepts for describing the micro-climatic behavior in urban areas. Sensors deployed for measuring these air and building characteristics are stationary.

For the second domain, water quality indicators (e.g., pH) and water characteristics (e.g., temperature and current speed) are the most important concepts. However, significant attention has been paid to the navigational concepts as the readings of these water quality indicators are often taken by sensors mounted on autonomous vehicles. Location information for the individual vehicles or groups of vehicles as well as the location for the potential danger zones (e.g., zones with underwater barriers) are thus important for researchers to make sense of the data collected.

To capture these domain concepts and to support intelligent search across domains, the ontology design adopts a strategy that is in line with the recommendations from the ontology development community – that the ontology is modular [24], has a clearly delineated content [25], is based on a well-designed upper-level ontology [26], and is independent from any databases [11]. The strategy leads to a two-layer ontology that consists of two domain ontologies at the bottom, and cyberinfrastructure (CI) ontology at the top. CI ontology acts as the basis of domain ontology. Concepts in the CI ontology such as *Space* and *Time* are generic to both domain ontologies and are useful for defining domain concepts in a more consistent manner.

Some of the concepts in the SWEET Ontology are adopted in the CI ontology as it provides a common semantic framework for earth science domains. The SWEET concepts such as *NumericalEntity*, *PhysicalProperty*, *Instrument*, *HumanActivity*, and *Unit* were chosen to be the core CI concepts of the following seven CI main categories (Figure 1) due to their relevancy to the domains in question (note that every category includes the related concepts as well as the core concepts indicated by its name) :

1. **Data**. Data is the core concept of the CI ontology and has a much richer meaning than most other concepts because it can be instances of any others. It is also connected to many other concepts, including data accessing forms (e.g., *DataFile* and *DataService*), data format (e.g., *Text* and *Binary*), data attribute (e.g., *Size* and *Format*), and spatial data model (e.g., *Vector* and *Raster*). It is mainly related to the Data in SWEET and the concepts and relations from CSDGM.

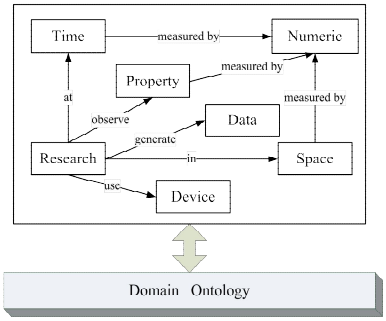

Figure 1. The Main Concepts of CI Ontology

2. **Property**. The concept describes physical and spatial quality associated with an object. For example, physical properties such as *Temperature*, *Weight*, and *Length*, are applicable for most physical objects. Spatial properties such as *Location*, *Orientation*, and *Elevation*, are applicable for objects which are in a space coordinate system. It is mainly related to *PhysicalProperty* of SWEET.
3. **Device.** This concept describes all the hardware used in the research. The most typical ones are *Computer*, *Sensor*, *Vehicle*, and *GPS*. Most of these concepts are sub-concepts of *sweet*:*Instrument*. Since the *sciInstrument* of SWEET has limited sub-concepts, concepts from SensorML and our research domains have been added (e.g., environment and geography).
4. **Research.** This concept incorporate any research domains, research actions (e.g., *Observation* and *Analysis Fieldwork*), and academic activities (e.g., *Conference* and *Publication*). It is mainly related to the *HumanActivity* of SWEET.
5. **Space.** This concept describes the basic characteristics of physical spaces of an object. An object can be associated with one or more two- or three-dimension properties that indicate its geometric characteristics such as location and shape. It also defines the basic frames and reference for spatial objects, including topology relations such as *containment* and *located-in*. *Space* is the basic category of SWEET, and its related *spaceCoordinates*, *spaceDirection*, *spaceDistribution*, and *spaceObject*.
6. **Time.** Temporal concepts are most common in any environmental data. Similar to spatial reference system, time is always associated with a reference system, such as Before Christ (B.C.) and Anno Domini (A.D.). Some computer systems may use other reference systems or their own customized systems. These concepts are related to the *Time* concept in SWEET. There are different units and reference systems for time, which are defined in ontology to assist the processing of temporal data.
7. **Numeric.** Numeric concepts are used to represent the quantity observed for a property. They are associated with values, units and reference systems, which are defined here and reused for specific subclasses. It is mainly related to *NumericEntity* and *Unit* of SWEET.

 ISBN: 978-1-61208-104-5

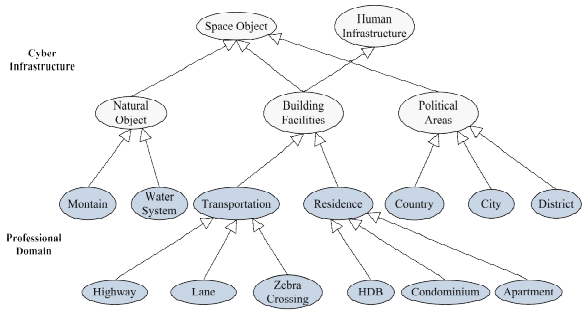

Figure 2. Concepts from different layers of ontology. HDB (Housing Development Bureau Flat, flats built by Housing Development Bureau of Singapore)

The concepts extracted from the domain are included in the domain ontologies. Each of these concepts holds an *is-a* relationship to one or more concepts in the CI ontology. For example, the *Residence* in the domain ontology is a *BuildingFacility* in the CI ontology (Figure 2). Concepts from SWEET are incorporated in domain ontology if they are deemed equivalent to the domain concepts, e.g., *Temperature*, *Humidity*, and *WaterPressure*.

The domain ontologies are enriched with the concepts from CSDGM and SensorML. The CSDGM defines the necessary metadata for a dataset, some of which have relations to *Data*, such as name, URI, spatial domain, spatial reference, size, and suffix, while others are sub-concepts, such as *ShapeFile* as a sub-concept of *DataFile*. SensorML provides a list of sensor concepts for us such as *Thermometer*, *Anemometer*, and *Barometer*. These concepts are all sub-concepts of *sweet:Instrument*. Furthermore, we specify the concepts such as *Input* and *Output*, which could be referred to *PhysicalProperty*, such as *WindPressure*, *WaterTemperature*.

From the development of the two-layered ontology several points were learned. First, existing metadata standards and upper-level ontologies such as SWEET generally contain concepts sufficient for describing the semantics of the environmental data. What is needed, however, is a clear distinction between concepts, relations, and a more comprehensive encoding of the relationships between these concepts. They enable the system to automatically identify and match the related concepts.

Second, the layered ontology is flexible for queries with different granularities. A user of the system can readily query related data and refine the query conditions with the help of the ontology. For example, to identify the temperature data, a user of the system might start with a search based on the concept *Temperature*, and then filter the result by its relations, e.g., spatial and time scope. The user might want to know the usability of this data, which might be met by giving them the sensor information from which the data are generated. Well-formed upper-level concepts and relations between them can be utilized to query the data by both domain and computational concepts.

Third, for the data engineers who are responsible to help both the data providers to publish their data and the data users to find the right data, the need to bridge the gap between domain concepts (e.g., temperature) and computational concepts (e.g., data service, file repository and database) cannot be overlooked. The task requires ontology to recognize computational concepts but not mix them up with the professional domain concepts. Separating the domain concepts in the domain ontology from computational concepts in the CI ontology helps maintain such conceptual clarity. In addition, it utilizes concepts in different ways, alleviating data engineers or providers from doing the conversion between the user interfaces and different cyber components.

IV. REASONING

Reasoning enables multiple interpretations of one or more basic concepts [27]. It also reduces the number of concepts that are left undefined while making precise the semantics of other concepts. In our work, reasoning is supported by three types of information: (1) explicitly declared relations between concepts, (2) T-Box axioms, and (3) rules. The explicitly declared relations, such as is-a, permits the reasoning of related concepts based on the axioms defined with the relations. T-Box axioms can be necessary or equivalent axioms that are used to infer new relations for existing data. The axioms can also be used to infer concepts associated with the concept in question and the relations between them. Such inference mechanism enables validation of the completeness of the data. Rules are specified by the ontology designer to indicate the implications between two sets of statements.

Examples of T-Box axioms are shown in Table 1. Axiom 1 validates if a metadata has provided the basic provenance information, which comes from either observation or process. Axiom 2 validates if a vector instance (e.g., time point) has been assigned reference information, e.g., UTC to a temporal value. Axiom 3 validates the completeness of a process definition. Axiom 4, 5, and 6 populate new concepts using instance from other concepts that make more sense to users from different professional domains.

Table 1. Examples of T-Box axioms**.** The "*some*" means there is at least one value coming from the range defined thereafter. The "*min*", "*max*" and "*exactly*" are cardinality constraints on the binary relations.

T-Box Axioms
1. Data ⊂ {has_source *some* Observation ∪ has_source *some* Process}
2. Vector ⊂ {Numeric ∩ has_reference *exactly* 1 }
3. Process ⊂ {has_input min 1 ∩ has_output *min* 1 ∩ has_processor *min* 1}
4. GeographicalData ≡ {Data ∩ has_model *some* SpatialDataModel}
5. ElevationData ≡ {Data ∩ (has_model *some* DEM ∪ has_model *some* DTM ∪ has_model some DSM ∪ has_model *some* Contour)}
6. Thermometer ≡ {Sensor ∩ has_input *some* Temperature }

Table 2. Rules for the reasoning on data. The "r(x,y)" means that binary relation r has the subject x and the object y. The "ist(x,y)" means that x is an instance of y. The "sub(x,y)" means x is a subclass of y. The "sup(x,y)" means x is a super class of y. The "eql(x, y)" means x equals to y.

	Rule
1.	has_content(?x,?c1) ∩ has_content (?y,?c2) ∩ (sub(?c1,?c2) ∪ sup(?c1,?c2) ∪ eql(?c1,?c2)) → compatible_content(?x,?y)
2.	ist(?p,Process) ∩ has_input(?p,?x) ∩ has_output(?p,?y)→ has_parent(?x,?y)
3.	ist(?x,TemperatureUnit) ∩ ist(?y,TemperatureUnit)→ convertible_unit(?x,?y)
4.	ist(GeoReference,?x) ∩ ist(GeoReference,?y)→ convertible_reference(?x,?y)
5.	ist(?x,Contour) ∩ ist(?y,DEM)→convertible_model(?x,?y) ist(?x,DEM) ∩ ist(?y,TIN)→convertible_model(?x,?y) ist(?x,DLG) ∩ ist(?y,DLG) ∩ has_feature_type(?x,?f) ∩ has_feature_type(?y,?f)→convertible_model(?x,?y)
6.	ist(?x, Data) ∩ generatedBy(?x, ?s) ∩ has_location(?s, ?p) → located_in(?x, ?p)

A portion of the rules which had been useful in supporting reasoning on the datasets and sensors in our work is shown in Table 2. Rule 1 is used to decide if two datasets contains the same domain concepts, which indicates the compatibility of the datasets. Rule 2 is used to infer the provenance relation between two data sets. Rule 3 indicates that units under the same category are compatible and amenable to conversion, which is useful to deciding if two numeric instances with the specified units are convertible. Rule 4 indicates whether geo-reference systems are convertible to each other, e.g., a local coordinate system without geo-reference components such as datum, projection, is not convertible to a geo-reference system. Rule 5 indicates four pairs of model which are considered as compatible. Rule 6 makes the data generated by the sensor inherit some relations from it. In this case, the location of the sensor is taken as the location of the data.

All data in our project can be the input of the reasoning engine. Figure 3 shows the correspondence between the CSDGM metadata and ontology concepts. The reasoning engine does the conversion by parsing the CSDGM metadata entries and creating instances of its corresponding ontology concepts, e.g., instances of *DataFile*. The reasoning engine also performs validation and inference during conversion by using all axioms and rules, which include the declared "is-a" relations and those from the above two tables.

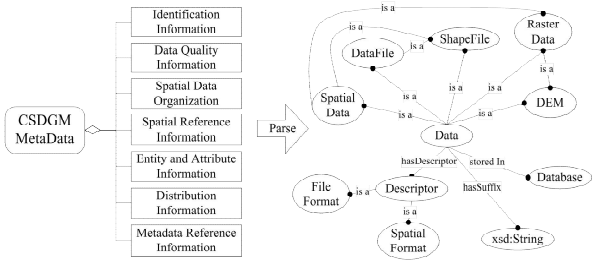

Figure 3. Parse the CSDGM metadata to ontology structure.

In our work, semantic gap exists between users of different domains. Reasoning function can be used to search the result with high precision and recall with the criteria which is not specified in the original query but stated either explicitly or implicitly in the ontology. Moreover, for the data engineers, reasoning function helps them to efficiently develop the program. Since axioms can be updated on the fly, costs of updating the programs in an ever-changing project can be sharply reduced.

V. ALIGNMENT AND TRANSLATION

Ontology alignment is the process of establishing correspondence between two similar concepts, including their subordinate and related concepts. In a data-centered scientific research, an alignment mechanism is needed to extract information from the data models of the original data sources, to perform reasoning, and to translate an ontology query request to a specific query language. A tool to support such process is vital in our work because the data producers and users often use different terms to refer to the same concepts and different encoding methods for their data.

Three types of alignment between commonly used data models and our domain ontologies were explored. Using the alignment between the CSDGM metadata and our domain ontologies (Figure 4), they include:

1. **Concept alignment** that identifies the corresponding concepts by text comparison. For example, the keyword *Depth* in an AUV dataset is identified as *Depth* (*of Water*). This alignment process is usually facilitated by referring to the context, i.e., the standard vocabulary or the ontology, used by the users.
2. **Instance alignment** that identifies the correspondence between instances in an ontology concept and semantic information in the database. These instances are typically extracted from rows in a database table or identified by unique reference identifiers.
3. **Relation alignment** that identifies the attribute of a data model to be a relation of an ontology concept, such as a temperature value of a "Temperature" concept, and other concepts essential to define the attribute, such as the unit for temperature.

One benefit of establishing these alignments is to facilitate the conversion of heterogeneous data models into a global ontological model preferred by the users of a particular domain [28]. Consider the following query: retrieve the climate data of Asia and return the records which were produced between *2009-3-3* and *2009-4-3*. To accomplish this task, the first step is to infer the possible candidates which meet the semantic requirement. In this case, there are two semantic restrictions: *is-a Climate* and *located-in Asia*. The relations *is-a* and *located-in* are both transitive relations. The reasoning process returns a list of candidate datasets related to *Climate*, such as *Temperature*, *Humidity*, and are located in *Asia*, e.g., *Japan* and *Singapore*. To further narrow in on the data in a particular time frame, assuming the data are stored in relational tables, the query with a time filter will be sent to the mediator for the relational tables and translated to the following:

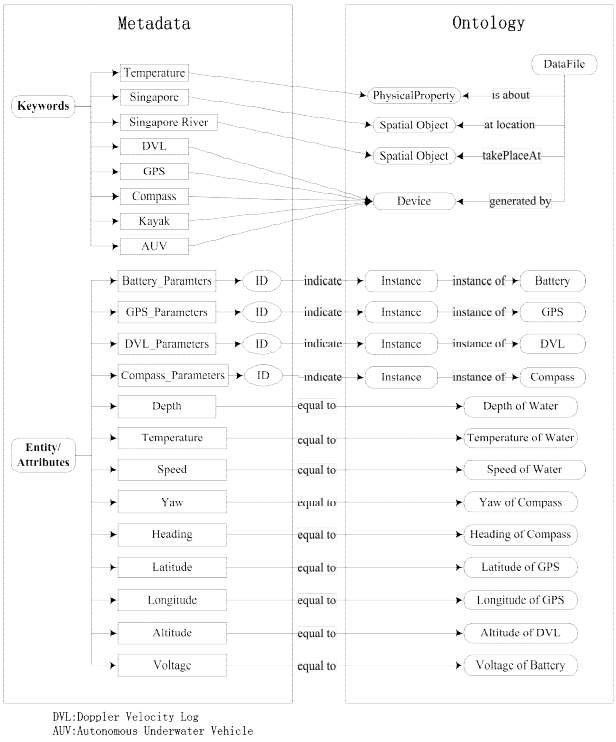

Figure 4. An example of the alignment between the metadata and the ontology.

Select value1, time1 from table1 where time1 > '2009-3-3' and time1 <'2009-4-3'

Select value2, time2 from table2 where time2 > '2009-3-3' and time2 <'2009-4-3'

Select value3, time3 from table3 where time3 > 1236009600000 and time3 <1238688000000

The last clause uses integer values for time representation (milliseconds between 1970-1-1 00:00:00 GMT and this time), which is used by some systems and needs to be translated to a uniform format. Alignment can take advantage of the *is-a* relation encoded in ontology concepts, i.e., if an alignment is applicable to an ontology concept, it could be applied to all its sub-concepts as well.

A second benefit of establishing these alignments is to support automatic conversion between scalars, between vectors, and between data formats:

1. **Scalar conversion**. This is a conversion for the data values described by a single scalar and an associated unit, such as different unit for Length, Area, Time, and Pressure.
2. **Vector conversion**. This is a conversion between data whose values are referenced to a chosen reference system. Spatial coordinates and time are typical examples of a vector.
3. **Text format conversion**. This is a conversion between different representation formats. For example, the text format for *Date* and *Time* varies in different data, even if they use the same unit and the reference system.

In our system, we focus on the alignment between ontology and conceptual or systematic data model. It is a process similar to ontology alignment except that a common data model is always vague on semantics, i.e., different types of entities and relations between them, where entity here can be regarded as an instance of a specified concept. Thus, the first step is to rebuild the entities and relations of the data model. In our system, we utilize ORM (Object-Relational Mapping) tool to achieve this on relational databases.

VI. IMPLEMENTATION

The system is a web application based on J2EE. We use Java as the major implementation language because it is widely supported by the open sources communities. We selected two ontology projects – Jena [28] and Pellet [29] – to process the ontology files. The OWL files are firstly generated by Protégé, and then stored and maintained via Jena API. The Pellet is a reasoning engine that provides support for SWRL (Semantic Web Rule Language [30]) based rules. We use ArcGIS from ESRI to develop spatial-related functions. The system architecture is shown in Figure 5.

The system adopts a three-layer architecture – the UI and Application Layer, Data Registration Layer, and the Resource Layer. The Data Registration layer uses the ontologies and other APIs to process the original information, which contains two paths, one for the data users and the other for the data providers:

For a data user, it is easier and more straightforward to search by concepts rather than look into the details of data. The user therefore uses the ontology to develop a request, which will be processed by the reasoning engine in the system and attached with richer semantics (e.g., more concepts and restrictions). The original query will be translated to different forms suitable for querying different data sources with the ontology alignment service.

A data provider provides sufficient metadata based on uniform standards, such as an XML schema which can be used to standardize the format of metadata, and is useful for standard-dependent programs. The metadata could be directly referred to ontology concepts, or some other standard vocabulary like GCMD, whereby they would then be recognized and aligned automatically. The data provider can also add more alignments or alter existing ones manually. Both of the original metadata and generated alignment are stored in the database for future applications.

VII. CONCLUSION AND FUTURE WORK

We have used ontology to model data semantics and to help users unfamiliar with the data structure and semantics to find the data they look for. We demonstrated the advantage of ontology-based system over traditional metadata or standard data exchange systems in bridging the semantic gaps in a heterogeneous environment. Ontology reasoning is a powerful tool for generating or importing a new ontology and its concepts without modifying the data. It makes it possible to accommodate concepts across different domains and from different user groups. Ontology alignment acts as a middleware for integrating data from different sources.

 ISBN: 978-1-61208-104-5

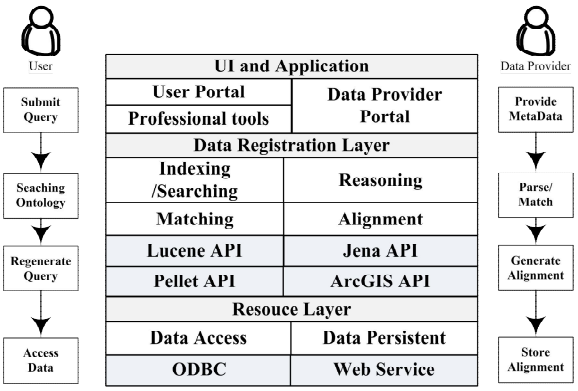

Figure 5. System architecture.

Compared to the most existing methods such as those mentioned in Section II, our work has focused on the following two points:

1. We have considered how to support data integration at both the semantic level and the conceptual model level. We give a clear roadmap from the users' request to the data retrieval, along with which ontology reasoning is essential for dataset searching and for integration.
2. We have designed and implemented a modular data integration system to ensure system flexibility. Every single application works independently while cooperating with each other through a dynamic, semantics-enabled interface. We also make sure the system is connected to the existing technologies and systems so that existing tools, e.g., metadata for populating databases and alignments for translating user queries to database queries, are reused.

Interfaces useful for automating data registration and alignment were developed. They allow a data set to be registered by uploading the associated metadata file compatible with CSDGM. They then automatically create instances of the metadata entries using the ontologies in the system. Alignments are mainly performed by the system managers who are well versed with the ontology concepts. Through text matching and ontology reasoning, the interfaces suggest the necessary inputs associated with the data that will be uploaded to the system.

Yet more remain to be incorporated to enrich the functionality of this system. First, with the increase of new applications and users, ontology is bound to evolve through time [31]. How to ensure the consistency of the whole ontology while evolving is an important problem to investigate. Second, extending the spatial reasoning capability of the system is crucial. For example, we can use spatial computation in the reasoning process to define the relations as *near*, *far*, and *neighbor*. Integrating spatial functions with ontology components particularly the reasoning functions would significantly improve the data search capability of the system. Third, users should be able to share both the data as well as the services associated with the formats of the data. For example, users could click a link to view a searched spatial dataset via an online visualization service. In this process, data should be organized and converted to a specific format suitable as the input of the service. The users can also choose to download them in a specific format and use local tools to handle them.

REFERENCES

[1] O. Dridi, "Ontology-Based Information Retrieval: Overview and New Proposition," Proc. 2nd International Conference on Research Challenges in Information Science (RCIS 2008), IEEE Press, Jun. 2008, pp. 421-426, doi: 10.1109/RCIS.2008.4632133.

[2] B. Barkallah and S. Moalla, "Metadata Driven Integration Model for Large Scale Data Integration," Proc. 7th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 09), IEEE Press, May. 2009, pp. 41-46, doi: 10.1109/AICCSA.2009.5069296.

[3] C. Youn, K. Kaiser, C. Santini and D. Seber, "Developing Metadata Services for Grid Enabling Scientific Applications," Proc. 6th International Conference on Computational Science (ICCS 2006), Springer-Verlag, May. 2006, pp. 379-386, doi: 10.1007/11758501_53

[4] A. Couchot, "Improving Web Searching Using Descriptive Graphs," Proc. Natural Language Processing and Information Systems (NLDB 2004), Springer-Verlag, Jun. 2004, pp. 276-287, doi: 10.1007/978-3-540-27779-8_24.

[5] N. H. Shah, C. Jonquet, A. P. Chiang, A. J. Butte, R. Chen and M. A. Musen, "Ontology-driven indexing of public datasets for translational bioinformatics," BMC Bioinformatics, vol. 10, Feb. 2009, doi: 10.1186/1471-2105-10-S2-S1.

[6] F. Fonseca and M. A. Rodriguez, "From geo-pragmatics to derivation ontologies: New directions for the geospatial semantic web," Trans. GIS, vol. 11, 2007, pp. 313-316.

[7] J. S. Madin, S. Bowers, M. P. Schildhauer and M.B. Jones, "Advancing ecological research with ontologies," Trends Ecol. Evol., vol. 23, 2008, pp. 159-168.

[8] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari and L. Schneider, "Sweetening ontologies with DOLCE," Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, vol. 2473, Springer-Verlag, Oct. 2002, pp. 223-233, doi: 10.1007/3-540-45810-7_18.

[9] "Suggested Upper Merged Ontology (SUMO)," http://www.ontologyportal.org/, [accessed 17 December].

[10] "Basic Formal Ontology (BFO)," http://www.ifomis.org/bfo, [accessed 16 July 2010].

[11] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications," Knowl. Acquis., vol. 5, 1993, pp. 199-220.

 ISBN: 978-1-61208-104-5

[12] "Semantic Web for Earth and Environmental Terminology (SWEET)," http://www.jpl.nasa.gov/ontology/, [accessed 17 July 2010].

[13] R. G. Raskin and M. J. Pan, "Knowledge representation in the semantic web for earth and environmental terminology (SWEET)," Comput. Geosci., vol. 31, Nov. 2005, pp. 1119-1125.

[14] A. Tripathi and H. A. Babaie, "Developing a modular hydrogeology ontology by extending the SWEET upper-level ontologies," Comput. Geosci., vol. 34, Sep. 2008, pp. 1022-1033.

[15] B. Brodaric and F. Probst, "DOLCE Rocks: Integrating Geoscience Ontologies with DOLCE," Proc. 2008 AAAI Spring Symposium, American Association for Artificial Intelligence (AAAI 2008), Mar. 2008, pp. 3-8.

[16] B. Beran and M. Piasecki, "Engineering New Paths to Water Data," Comput. Geosci., vol. 35, no. 4, Apr. 2009, pp. 753-760.

[17] B. Ludäscher, A. Gupta, and M. E. Martone, "Model-based Mediation with Domain Maps," Proc. 17th IEEE International Conference on Data Engineering (ICDE 2001), IEEE Comput. Soc., Apr. 2001, pp. 81-90, doi: 10.1109/ICDE.2001.914816.

[18] H. Wache, T. Vogele, and U. Visser, "Ontology-Based Integration of Information - A Survey of Existing Approaches," Proc. 17th International Joint Conferences on Artificial Intelligence (IJCAI 2001), Morgan Kaufmann, 2001, pp. 108-117.

[19] A. Buccella, A. Cechich, and P. Fillottrani, "Ontology-driven geographic information integration: A survey of current approaches," Comput. Geosci., vol. 35, Apr. 2009, pp. 710-723.

[20] D. J. Russomanno, C. R. Kothari, and O. A. Thomas, "Building a Sensor Ontology: A Practical Approach Leveraging ISO and OGC Models," Proc. International Conference on Artificial Intelligence (ICAI 2005), CSREA Press, Jun. 2005, pp. 637-643.

[21] "OGC, Sensor Model Language (SensorML)," http://www.opengeospatial.org/standards/sensorml, [accessed 17 July 2010].

[22] M. Shankar, A. Sorokine, B. Bhaduri, D. Resseguie, S. Shekhar, and J. S. Yoo, "Spatio-Temporal Conceptual Schema Development for Wide-Area Sensor Networks," Geospatial Semantics, vol. 4853, Springer-Verlag, Nov. 2007, pp. 160-176, doi: 10.1007/978-3-540-76876-0_11.

[23] "FGDC, Content Standard for Digital Geospatial Metadata," http://www.fgdc.gov/metadata/csdgm/, [accessed 17 July 2010].

[24] R. Rector, "Modularization of Domain Ontologies Implemented in Description Logics and Related Formalisms including OWL," Proc. 2nd International Conference on Knowledge Capture (K-CAP 03), ACM, 2003, pp. 121-128, doi: 10.1145/945645.945664.

[25] "The Open Biological and Biomedical Ontologies: Current Principles," http://www.obofoundry.org/crit.shtml. [accessed 17 July 2010].

[26] N. Guarino, "Formal Ontology in Information Systems," Proc. 1st International Conference on Formal Ontology in Information Systems (FOIS 98), IOS Press, 1998, pp. 3-15.

[27] J. Z. Pan and I. Horrocks, "Web Ontology reasoning with Datatype Groups," Proc. 2nd International Semantic Web Conference (ISWC2003), Springer-Verlag, Oct. 2003, pp. 47-63, doi: 10.1007/978-3-540-39718-2_4.

[28] "Jena - A Semantic Web Framework for Java, " http://jena.sourceforge.net/, [accessed 17 July 2010].

[29] "Pellet: The Open Source OWL Reasoner, " http://clarkparsia.com/pellet/, [accessed 17 July 2010].

[30] "OGC, SWRL: A Semantic web Rule Language Combining OWL and RuleML, " http://www.w3.org/Submission/SWRL/, [accessed 17 July 2010].

[31] G. Flouris, D. Plexousakis, and G. Antoniou, "Evolving Ontology Evolution, " SOFSEM 2006: Theory and Practice of Computer Science, vol. 3831, Springer, Jan. 2006, pp.14-29, doi: 10.1007/11611257_2.

 ISBN: 978-1-61208-104-5